Alternatives to the Chi-Square Test for Evaluating Rank Histograms from Ensemble Forecasts

KIMBERLY L. ELMORE*

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

(Manuscript received 28 February 2005, in final form 20 May 2005)

ABSTRACT

Rank histograms are a commonly used tool for evaluating an ensemble forecasting system's performance. Because the sample size is finite, the rank histogram is subject to statistical fluctuations, so a goodness-of-fit (GOF) test is employed to determine if the rank histogram is uniform to within some statistical certainty. Most often, the χ^2 test is used to test whether the rank histogram is indistinguishable from a discrete uniform distribution. However, the χ^2 test is insensitive to order and so suffers from troubling deficiencies that may render it unsuitable for rank histogram evaluation. As shown by examples in this paper, more powerful tests, suitable for small sample sizes, and very sensitive to the particular deficiencies that appear in rank histograms are available from the order-dependent Cramér–von Mises family of statistics, in particular, the Watson and Anderson–Darling statistics.

1. Introduction

Rank histograms are used extensively to evaluate ensemble forecast system performance (e.g., Hamill and Colucci 1997, 1998; Hou et al. 2001; Stensrud and Yussouf 2003). Rank histograms were introduced into the climate field by Anderson (1996), and Anderson and Stern (1996) used the Kolmogorov–Smirnov test along with the Anderson–Darling test for comparing samples in seasonal simulation cases. Hamill (2001) shows how to appropriately use rank histograms for evaluating ensemble forecasts. Once biases in individual members are removed and observational error is accounted for (Hamill 2001), the ideal ensemble produces flat, or uniform, rank histograms. Certain deviations from the uniform distribution are bellwether indicators that the ensemble forecasting system is deficient. Depending on the nature of these deviations, the nature of the deficiency may be better defined. For example, a U-shaped distribution indicates the ensemble is underdispersive, a peaked distribution suggests that the ensemble is overdispersive, and a sloped rank histogram indicates that the ensemble remains biased in some way.

Due to random variations, even ideal ensembles will not produce perfectly uniform rank histograms. Hence, one wishes to test the assumption that, within sampling error, the rank histogram is derived from a discrete uniform distribution. Such tests are derived from the general family of goodness-of-fit (GOF) tests, which test the null hypothesis, H_0: the rank histogram is indistinguishable from a discrete uniform distribution. A common test for evaluating whether the resulting rank histograms come from a discrete uniform distribution is the χ^2 test. But the χ^2 test is far from ideal and lacks power for small sample sizes. More powerful tests come from the Cramér–von Mises (CvM) family of statistics, specifically the Watson test and the Anderson–Darling test, which are described in section 2. Section 3 discusses the results of applying the different tests to both large and small datasets generated by sampling at random from a uniform distribution. Section 4 provides conclusions and recommendations.

2. GOF tests

The most common GOF test is the χ^2 test. This is a natural test for rank histograms, which represent data binned by rank. The χ^2 test is defined as follows by the test statistic, *T:*

* Additional affiliation: NOAA/National Severe Storms Laboratory, Norman, Oklahoma.

Corresponding author address: Dr. Kimberly L. Elmore, NSSL, 1313 Halley Circle, Norman, OK 73069.
E-mail: kim.elmore@noaa.gov

$$T = \sum_{i=1}^{k} (O_i - F_i)^2/F_i, \quad (1)$$

where O_i is the observed frequency in bin i, and F_i is the expected frequency in bin i under the null distribution with k cells. The T statistic for the null distribution is approximately distributed as χ^2 with $k - 1$ degrees of freedom.

Sample size is always an issue with GOF tests. In practice, GOF tests have limited value for both very large, and very small, sample sizes, though what constitutes "very large" and "very small" is not usually clear and differs from test to test. If the sample size is large enough, almost any GOF test will reject the null hypothesis because real data are never distributed according to any theoretical distribution (Millard 2002). As the sample size decreases, the power, or ability to detect a difference between the sample distribution and the hypothesized or null distribution (uniform, in this case), of any test suffers, though certain tests are more sensitive than others against particular alternative hypotheses for any given sample size.

For the χ^2 test, the conservatively defined required sample size is that which would provide an expected count of at least 5 for each bin. Thus, for a 15-member ensemble, the resulting rank histogram has 16 bins and 90 cases are required. However, the χ^2 approximation for the T statistic remains valid for surprisingly small samples. If N is the number of samples, c is the number of bins, and E_i is the expected frequency in bin i under the null hypothesis, the T statistic is still well approximated by the χ^2 distribution with $c - 1$ degrees of freedom if $N \geq 10$, $c \geq 3$, $N^2/c \geq 10$, and $E_i \geq 0.25$, which means that for a 15-member ensemble the minimum number of cases must be no less than 13. Conover (1999) gives a good treatment of how to compute the required sample size for a χ^2 test.

Other useful GOF tests exist. A notable example is the Cramér–von Mises (CvM) family of test, which has forms for the discrete uniform distribution (Choulakian, et al., 1994). This family consists of the Cramér–von Mises (Cramér 1928; von Mises 1931; Smirnov 1936), the Watson (Watson 1961), and the Anderson–Darling (Anderson and Darling 1952) tests. In general, the CvM family of GOF tests has more power than does the χ^2 test for small sample sizes. Unlike the χ^2 test, the CvM test statistics are nonparametric. The CvM provides nearly identical results as the Kolmogorov–Smirnov (KS) test, though some find its formulation more appealing because the CvM tests use an integrated departure of the empirical distribution function (EDF) from the null distribution, instead of the largest departure (Conover 1999).

Consider a discrete distribution with k cells with probability p_i of an observation landing in any cell. Let o_i be the observed number of counts in bin i, and let, $Np_i = e_i$ be the expected number of counts in bin i under the null distribution. Then, let $S_j = \Sigma_{i=1}^{j} o_i$, and $T_j = \Sigma_{i=1}^{j} e_i$. Thus, S_j/N and $H_j = T_j/N$ correspond to the EDF $F_N(x)$. Finally, let $Z_j = S_j - T_j$, $j = 1, 2, \ldots, k$. Then the discrete form of the CvM statistic is given by

$$W^2 = N^{-1} \sum_{i=1}^{k} Z_j^2 p_j, \quad (2)$$

the discrete form of the Watson statistic is given by

$$U^2 = N^{-1} \sum_{i=1}^{k} (Z_j - \overline{Z})^2 p_j, \quad (3)$$

and the discrete form of the Anderson–Darling statistic is given by

$$A^2 = N^{-1} \sum_{j=1}^{k} Z_j^2 p_j/[H_j(1 - H_j)], \quad (4)$$

where $\overline{Z} = \Sigma_{j=1}^{k} Z_j p_j$.

By definition, $Z_k = 0$, so the last term in $W^2 = 0$, and the last term in $A^2 = 0/0$, which is set to 0. An alternative notation is to extend the index over which the sums operate to only $k - 1$ in (2) and (4) (Choulakian et al. 1994).

Because the CvM family uses an integrated departure from the EDF, it is order dependent, which means the way the bins are indexed affects the value of the computed statistic. For example, the discrete CvM test statistic has the following form: $W^2 \sim (O_1 - F_1)^2 + (O_1 + O_2 - F_1 + F_2)^2 + (O_1 + O_2 + O_3 - F_1 + F_2 + F_3)^2 + \ldots$, and so the order in which the binned values appear affects the statistic's value. This is also true for the Anderson–Darling statistic. This is only partially true for the Watson statistic, which has a circular dependence. The Watson statistic differs from the other two in that it is invariant with regard to the "starting" cell. Regardless of the start index in (3), as long as all indices are addressed in order thereafter, the Watson statistic is invariant for any given dataset. Hence, the Watson statistic is particularly useful for testing the uniformity of counts around a circle (Choulakian et al. 1994), such as calendar months or wind direction. All three, however, apply to linear data. The distribution theory, and so methods for constructing p values for the CvM, Watson, and Anderson–Darling statistics, are all given in Choulakian et al. (1994) and will not be elaborated upon here.

In contrast, the χ^2 test is insensitive to the nature of the departure from the null distribution because it uses

TABLE 1. Significant values for Cramér–von Mises statistics for tests of the discrete uniform distribution with k cells, α = upper-tail significance test.

Cramér–von Mises statistic, W^2								
k	$\alpha = 0.25$	0.15	0.10	0.05	0.025	0.01	0.005	0.001
3	0.1980	0.282	0.351	0.472	0.603	0.783	0.922	1.215
4	0.205	0.284	0.351	0.470	0.595	0.767	0.899	1.213
5	0.207	0.284	0.350	0.467	0.590	0.750	0.888	1.204
6	0.208	0.284	0.349	0.465	0.587	0.754	0.883	1.188
8	0.209	0.284	0.348	0.464	0.584	0.749	0.877	1.179
10	0.209	0.284	0.348	0.463	0.583	0.748	0.874	1.175
20	0.209	0.284	0.347	0.462	0.581	0.743	0.871	1.170
40	0.209	0.284	0.347	0.461	0.581	0.744	0.870	1.168
∞	0.209	0.284	0.347	0.461	0.581	0.743	0.869	1.167
Watson statistic, U^2								
k	$\alpha = 0.25$	0.15	0.10	0.05	0.025	0.01	0.005	0.001
3	0.103	0.141	0.171	0.222	0.273	0.341	0.395	0.512
4	0.106	0.139	0.165	0.209	0.252	0.309	0.351	0.453
5	0.107	0.137	0.161	0.201	0.241	0.294	0.335	0.427
6	0.107	0.136	0.158	0.197	0.235	0.286	0.325	0.414
8	0.106	0.134	0.156	0.193	0.230	0.278	0.315	0.401
10	0.106	0.133	0.154	0.191	0.227	0.275	0.311	0.395
20	0.105	0.132	0.152	0.188	0.223	0.270	0.305	0.388
40	0.105	0.131	0.152	0.187	0.222	0.269	0.304	0.387
∞	0.105	0.131	0.152	0.187	0.222	0.268	0.304	0.385
Anderson–Darling statistic, A^2								
k	$\alpha = 0.25$	0.15	0.10	0.05	0.025	0.01	0.005	0.001
3	0.892	1.267	1.580	2.125	2.714	3.52	4.15	5.47
4	0.989	1.363	1.675	2.235	2.821	3.63	4.24	5.71
5	1.043	1.417	1.733	2.289	2.874	3.68	4.28	5.77
6	1.079	1.452	1.763	2.324	2.909	3.72	4.33	5.80
8	1.122	1.495	1.807	2.367	2.952	3.72	4.37	5.84
10	1.147	1.521	1.832	2.392	2.977	3.78	4.40	5.88
∞	1.248	1.621	1.933	2.492	3.077	3.88	4.50	5.97

only the sum of the individual deviations at each bin over all bins. Hence, the χ^2 test cannot distinguish between noisy departures from the null distribution and U-shaped, peaked, or sloped departures, which are ordered departures. The CvM family is relatively insensitive to random departures but more sensitive to ordered departures from the null distribution, and retains more power against these departures than does the χ^2 test. This can lead to profound differences between GOF test results based on the χ^2 test and results based on the CvM statistics. Table 1 shows critical values for the various significance levels for the three CvM statistics (from Choulakian et al. 1994).

3. Examples

Differences between the χ^2 and CvM test behavior are illustrated using a Monte Carlo simulation that draws random samples from a uniform distribution. Assume an ensemble forecasting system with 15 members, which yields a rank histogram containing 16 bins. Define a small-sample rank histogram as consisting of 60 cases. Thus, for the small-sample simulation, each of 1000 Monte Carlo trials draws 60 numbers uniformly distributed between 1 and 16 (Fig. 1). Define a large-sample rank histogram as consisting of 540 cases. So, the large-sample Monte Carlo simulation uses 1000 sets of 540 numbers uniformly distributed between 1 and 16 (Fig. 2). In the small-sample case, the expected number of counts, E_i, in each cell is 3.75, large enough for both the χ^2 test and the CvM family of tests to be valid. For the large sample, $E_i = 33.75$. Each sample is then reordered pathologically to produce a U-shaped, a peaked-shaped, and a sloped bias trend. Hence, each sample results in four possible distributions: random, U shaped, peaked, and sloped. In each case, the χ^2 test p value remains invariant, but the CvM tests vary widely depending on the nature of the reordering.

For a test at $p = 0.05$, the expectation is that close to 5% of the samples will result in a p value less than 0.05 for all of these tests simply by random chance before the samples are reordered. By definition, the χ^2 test p value is independent of order. However, the CvM family yields different p values, depending on how the data are ordered. Table 2 shows the proportion of cases that

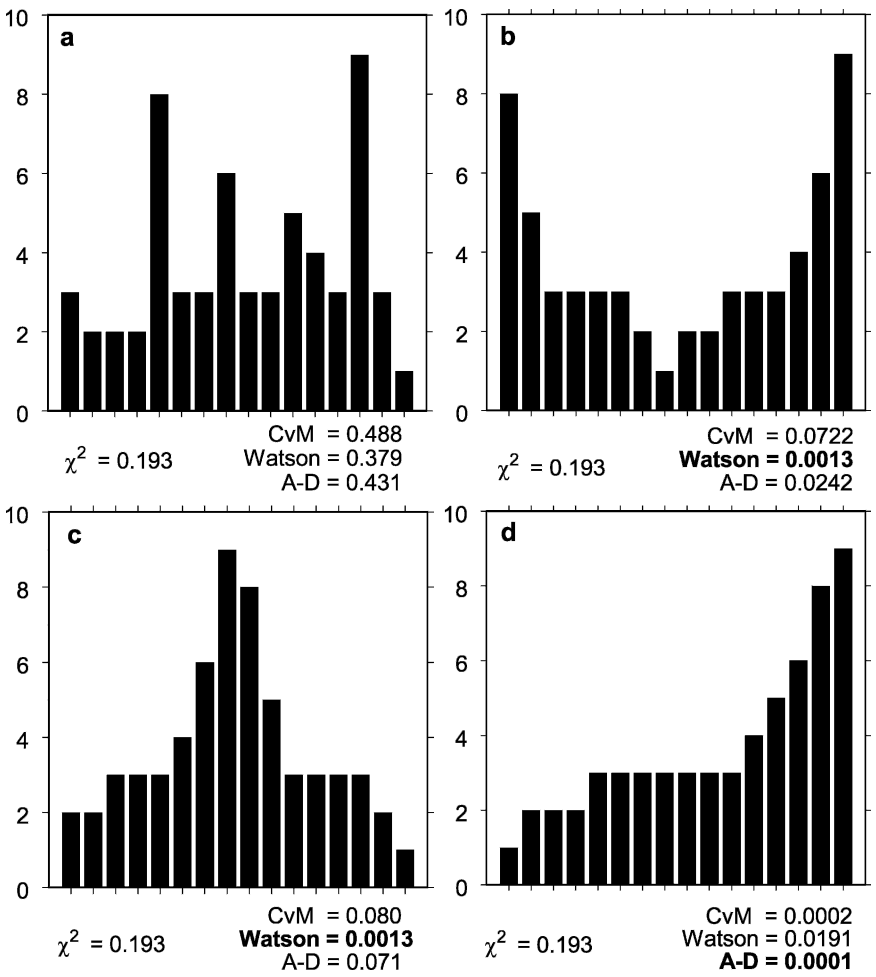

FIG. 1. Examples of χ^2 and Cramér–von Mises family test results for a single, 60-element dataset with different bin ordering. The *y* axis yields the number of elements in each bin, and the *x* axis is the rank, with rank 1 on the left and rank 16 on the right. Tests that are particularly sensitive to certain deviations from the uniform null distribution are in boldface. (a) Rank histogram of uniformly distributed data with noise, (b) same data as in (a) but with the data reordered to generate a U-shaped rank histogram, (c) same data as in (a) but with data reordered to create a peaked rank histogram, and (d) same data as in (a) but reordered to generate a sloping rank histogram.

are associated with p values ≤ 0.05 for each ordering for the small sample size, while Table 3 shows the same results for the large sample size (boldface values are associated with the most sensitive tests).

Clearly, for both sample sizes, the expectation for the random rank histograms is met by all tests. However, for the pathologically reordered rank histograms, the statistics show marked differences. While the χ^2 statistic is unaffected, the Watson statistic is clearly very sensitive to U-shaped and peaked rank histograms (yielding identical values for both due to circular symmetry), while both the CvM and Anderson–Darling tests are most sensitive to the bias slope rank histograms. These results hold for both the small and large samples.

More insight may be gained by examining how the CvM tests behave relative to the χ^2 test (Fig. 3). For the random data, the two GOF tests are uncorrelated for 16-bin rank histograms constructed from both 60 cases (Fig. 3a) and 540 cases (Fig. 3c), which means that the individual samples with p values <0.05 differ from the two tests in nearly random ways. However, the χ^2 test is by definition insensitive for the data reordered into a U shape, while the Watson statistic is clearly very sensitive to this reordering for both the 60- (Fig. 3b) and 540-case samples (Fig. 3d). The χ^2 test p values are clearly discrete for the small-sample (60 values) cases (Figs. 3a and 3b), which is a result of both the small sample size and the insensitivity to order: with only 60

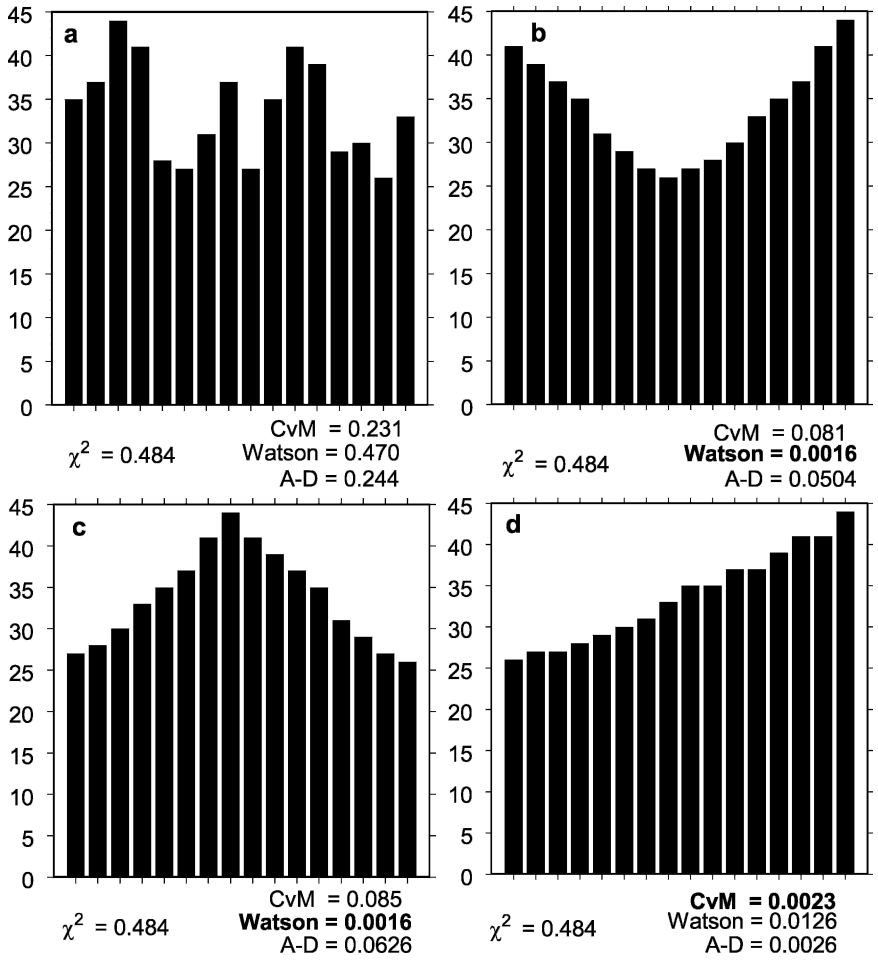

FIG. 2. Same as in Fig. 1 but for a sample size of 540.

values, the number of different *T*-statistic values that can be generated is significantly limited. The Watson test *p* value rises as the χ^2 test *p* value approaches 1 (Figs. 3b and 3d). This is true for all of the CvM family of tests, and indicates that as the χ^2 test *p* value approaches 1, the rank histogram approaches exact uniformity. Because the sample size used in the simulations is not a multiple of the number of bins, such a "perfect" rank histogram cannot occur within these data. However, were all bins to contain the same number, the rank histogram itself is invariant under bin-order permutations. The sensitivity of the Watson statistic to U-shaped distributions is demonstrated by how high the χ^2 test *p* value may become before the Watson test *p* value begins to increase (Figs. 3b and 3d). The higher the χ^2 test *p* value at which the Watson *p* value starts to rise, the more sensitive the Watson test statistic is to the particular deviation from the null distribution. Simulations that use sample sizes as large as 5000 and as small as 30 show similar results.

4. Conclusions

The most common problems associated with ensemble forecasting systems are underdispersion, over-

TABLE 2. Small-sample results. Bold values are associated with the most sensitive tests.

	Random	Reordered: U shaped	Reordered: peak shaped	Reordered: bias slope
χ^2	0.055	0.055	0.055	0.055
CvM	0.057	0.223	0.183	**0.995**
Watson	0.051	**0.923**	**0.923**	0.583
Anderson–Darling	0.057	0.451	0.325	**0.994**

TABLE 3. Large-sample results. Bold values are associated with the most sensitive tests.

	Random	Reordered: U shaped	Reordered: peak shaped	Reordered: bias slope
χ^2	0.055	0.055	0.055	0.055
CvM	0.042	0.228	0.223	**0.994**
Watson	0.052	**0.911**	**0.911**	0.580
Anderson–Darling	0.042	0.454	0.405	**0.994**

dispersion, and bias. Rank histograms depict these problems with either a U shape, a peaked center, or one end of the rank histogram being higher than the other (slope), respectively. The ubiquitous χ^2 test possesses certain characteristics that may make it unsuitable for assessing the quality of rank histograms derived from ensemble forecasting systems. Specifically, the χ^2 test statistic is order invariant, which means rank histograms that clearly display a problem within the ensemble forecasting system may go undetected using the χ^2 test.

Better tests for the specific problems encountered with ensemble forecasting systems are available from the discrete form of the Cramér–von Mises family of GOF tests. These tests are based on the Cramér–von Mises, Watson, and Anderson–Darling statistics. Of these three, the Watson test statistic is considerably more sensitive to either U-shaped or peaked rank his-

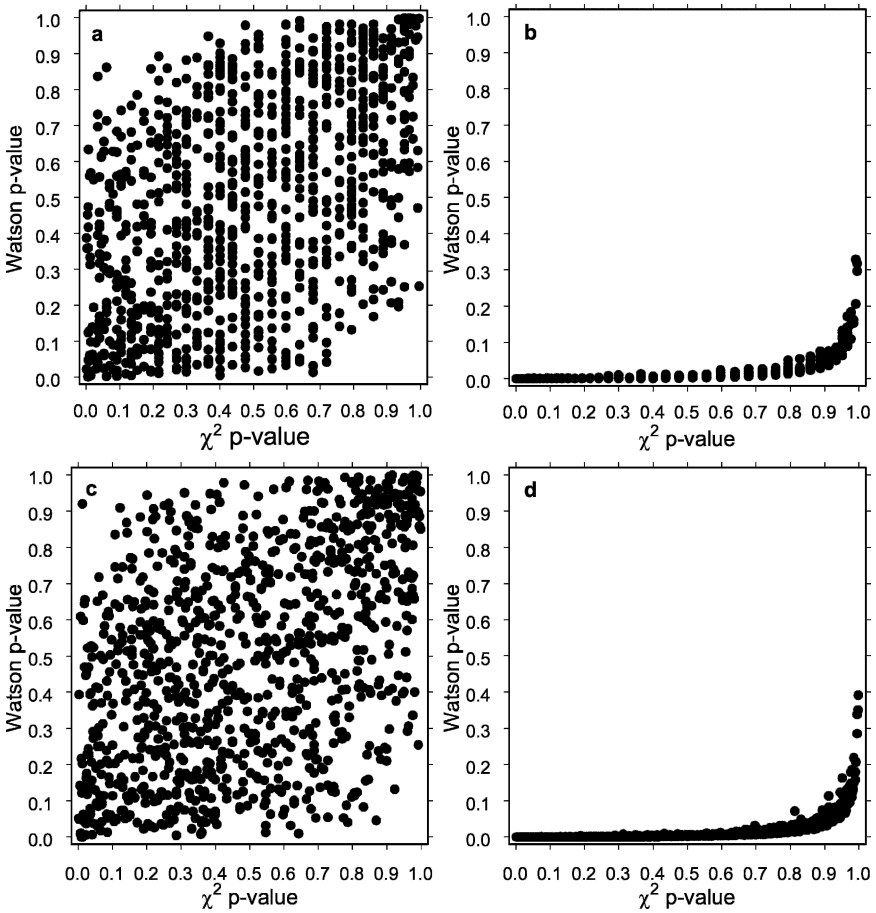

FIG. 3. Examples of the χ^2 test behavior compared to the Watson test behavior: (a) plot showing the association between the χ^2 p value and the Watson p value for the random data consisting of 60 values, (b) χ^2 p value and the Watson p value for the reordered into a U shape for 60-sample data, (c) same as in (a) but for the 540-sample data, and (d) same as in (b) but for the 540-sample reordered data.

tograms than are the other two. The other two tests are almost equally sensitive to a slope/bias within the rank histograms. All of the CvM tests retain considerable power for relatively small samples.

Like any statistical test, the CvM tests are not infallible. Yet, better results will be obtained by using a combination of either the Watson and CvM, or Watson and Anderson–Darling tests for evaluating rank histograms, instead of the χ^2 test. If the rank histogram in question passes either combination of CvM tests at an appropriate p value, then that rank histograms may be considered statistically free from either U-shaped/peaked or biased/sloped deficiencies.

Acknowledgments. The author is very grateful to Dr. Richard A. Lockhart, of Simon Fraser University, Burnaby, British Columbia, Canada, without whose generous guidance this work would not have been possible. This work is supported by the National Severe Storms Laboratory.

REFERENCES

Anderson, J. S., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integration. *J. Climate,* **9,** 1518–1530.

——, and W. F. Stern, 1996: Evaluating the potential predictive utility of ensemble forecasts. *J. Climate,* **9,** 260–269.

Anderson, T. W., and D. A. Darling, 1952: Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.,* **23,** 193–212.

Choulakian, V., R. A. Lockhart, and M. A. Stephens, 1994: Cramér–von Mises statistics for discrete distributions. *Can. J. Stat.,* **22,** 125–137.

Conover, W. J., 1999: *Practical Nonparametric Statistics.* 3d ed. John Wiley and Sons, 584 pp.

Cramér, H., 1928: On the composition of elementary errors. *Skand. Aktuarietidskr.,* **11,** 13–74, 141–180.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX'98 ensemble forecasts. *Mon. Wea. Rev.,* **129,** 73–91.

Millard, S. P., 2002: *Environmental Stats for S-Plus.* 2d ed. Springer, 264 pp.

Smirnov, N. V., 1936: Sui la distribution de w^2 (Criterium de M.R.v. Mises). *Compt. Rend.,* **202,** 449–452.

Stensrud, D. J., and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.,* **131,** 2510–2524.

von Mises, R., 1931: *Wahrscheinlichkeitsrechnung und Ihre Anwendung in der Statistik und Theroretishen Physik.* Vol. 1. F. Deuticke, 574 pp.

Watson, G. S., 1961: Goodness-of-fit tests on a circle. I. *Biometrika,* **48,** 109–114.